Detection of Cyberbullying in the Social Media Space using Maximum Entropy and Convolutional Neural Network

ABSTRACT

Social media has shown to be a medium for the exchange of information, conducting commerce, and even for religious and political activity. The same site has been used to disseminate cyberbullying and hate speech. Cyberbullying is an offense when a perpetrator targets a victim with online provocation and resentment, which has negative impacts on the victim's emotions, relationships, and physical health. These spew forth a lot of hate speech directed at those who have different beliefs or hold different opinions. In order to overcome these problems, this research creates a deeper neural network and maximum entropy-based model for the identification of cyberbullying. In comparison to the present systems, Convolution Neural Network is used for the superior results. The dataset gathered includes 24,783 records tweets with the categories "bullying language," "non-bullying language," and "neither" are included. According to geopolitical zones, tweets are categorized in the study. Throughout the experiment, the dataset was trained and evaluated. The model's unigram performed with 96.3 percent accuracy, the bigram model with 93.8 percent accuracy, the trigram with 88.2 percent accuracy, and the n-gram enhanced performance with 94.2 percent. The two baseline classifiers for the tweets dataset were TF-IDF (Term Frequency Inverse Document Frequency) and the characteristics of unigram, bigram, trigram, and n-grams for detection. The outcome of the model during the tweet cyberbullying will increase the memory accuracy of the bullying. The study's findings will reveal creative language and alternative spellings for cyberbullying in tweets..

Keywords: ***Cyberbullying, Hate Speech, Maximum Entropy, Sentiment, Word Embedding***

1. INTRODUCTION

Hate speeches are statements, written materials, commercials, plays, musicals, or other works of literature that target a specific person or group of people based on their gender, race, religion, politics, or other factors. Cyberbullying is the term we use when describing this behavior when it occurs online in a social media setting. In certain nations, incitement to violence, sedition, and verbal abuse laws can apply to hate speech. According to [1], "Hate speech

includes any statement, gesture, behaviour, writing, or exhibition that might inspire individuals to commit acts of violence or prejudice." In essence, these speeches deprive other people of their dignity. According to the statement, "Hate speech uses derogatory terms to disparage and stigmatize individuals based on their race, ethnicity, gender, sexual orientation, or other characteristics of group membership. Any such words, actions, or behaviour without hate speech and the hatred it propagates, motivated individuals may attack any group with violence. Numerous studies on post-election violence have been undertaken, but little recent attention has been paid to evaluating the root causes of this violence in connection to hate speech and the role they play in fueling cyberbullying in global social media forums. Furthermore, it shouldn't be shocking that newspaper speech has been under scrutiny and still is given the influence and importance of news journalism in contemporary society. The study of people's views, sentiments, evaluations, appraisals, attitudes, and feelings regarding things including products, services, organizations, people, issues, events, and themes is known as sentiment analysis or opinion mining. evaluate mining, etc. Almost all human actions revolve around opinions since they have a significant impact on how we behave. We always want to hear what other people have to say before making a choice. In the real world, companies and organizations are constantly interested in learning what the general public or consumers think about their goods and services. Before making a purchase choice or casting their vote in a political election, individual consumers also want to hear what other people think of the products already in use and the political candidates they are considering. This study applies sentiment analysis to this area in an effort to identify and, if feasible, combat the hate speech that has a detrimental viral effect on cyberbullying as a result of social media platforms.

2. MATERIAL AND METHODS

One of the most active study fields in natural language processing (NLP) today is sentiment analysis [3] [4]. It has several complex and connected subproblems, including the categorization of sentiment at the phrase level. Many academics came to the realization that different phrase types require distinct approaches to sentiment analysis. For sentiment analysis, models of a variety of sentence kinds, such as subjective, target-dependent, comparison, negation, conditional, and sarcastic phrases, have been developed. Sentences that reflect views are referred to as subjective sentences, whereas sentences that represent factual information are referred to as objective sentences[5]. Many academics see subjectivity and sentiment as the same idea, even though some objective statements might infer feelings or opinions and some subjective words may not reflect any opinions or sentiments [6]. subjective expression in sentences sentiment analysis at the idea level and twitter. Prior to computing the sentiments at the sentence level, the proposal in [9] suggested a hybrid technique employing Senti WordNet [10] and fuzzy sets to estimate the semantic orientation polarity and intensity of sentiment words. [11] introduced a lexicon-based sentiment categorization method that captures contextual polarity from both local and global context for social media genres[12,13]. [14] proposed an unique strategy based on an unsupervised dependency parsing-based text classification algorithm for predicting sentiment in online documents.

The majority of earlier target-related studies presupposed that targets had been provided before doing sentiment analysis [15]. [16] [17] [18]. Despite the fact that there is a considerable body of work concentrating on opinion target extraction from sentences, little study has been done on categorizing sentences by the target number. As features for a naive Bayes classifier, comparable phrases were recognized using class sequence rules based on human-compiled terms that indicated comparison. According to, their work was the first to mine views from comparison phrases. To evaluate whether the aspect and sentiment context were more closely related in Pros or in Cons, they used linguistic norms and a sizable external corpus of Pros and Cons drawn[19] from product evaluations. A corpus of comparative statements from English camera evaluations was offered in the review in two linguistic knowledge-driven techniques for the extraction of Chinese comparative components were proposed. Negative sentences appear in the sentiment analysis corpus rather frequently. Due to the removal of the symbol-by-symbol constraint, the research's established model will make sure security specialists are able to set up protection mechanisms to lessen the likelihood of such assaults when early detection of incoming threats is made possible [20].
There is a trade-off between the amount of configuration messages and the number of permanent flow entries in the network as a result of the configuration messages flooding the control plane.

A compositional model to identify valence shifters, such as negations, which aid in the understanding of the polarity and strength of opinion expressions was presented by several scholars who took this into consideration in. The effects of modifiers on the emotions impacted by negation, intensifiers, and modality were explored in. Another linguistic element that appears frequently in text are conditional phrases. Usually, such a phrase includes To assess if the attitudes stated on various themes in a conditional sentence are positive, negative, or neutral, linguistic analysis of conditional sentences and the construction of several supervised learning models were used. A list of fascinating conditional phrase patterns that frequently convey emotion was provided, which was especially helpful for product evaluations, online forums, and blogs. In online forums, sarcasm is a sophisticated speech act that is frequently employed. In the context of sentiment analysis, this indicates that people often misinterpret their own words and say one thing when they mean another. A brand-new semi-supervised system for sarcasm recognition that could detect sarcastic language in product evaluations was the center of the emphasis. Using this corpus, the report looked at the effects of creating a corpus of snarky Twitter messages.

3. EXPERIMENTAL DETAILS

The adoption of the sequence model is necessary for this research project because the Order-insensitive models are insufficient to fully capture the semantics of natural language because they are unable to account for differences in meaning resulting from differences in word order or syntactic structure (e.g., "cats climb trees" vs. "trees climb cats"). Figure 1 depicts the model that was created for this study project.

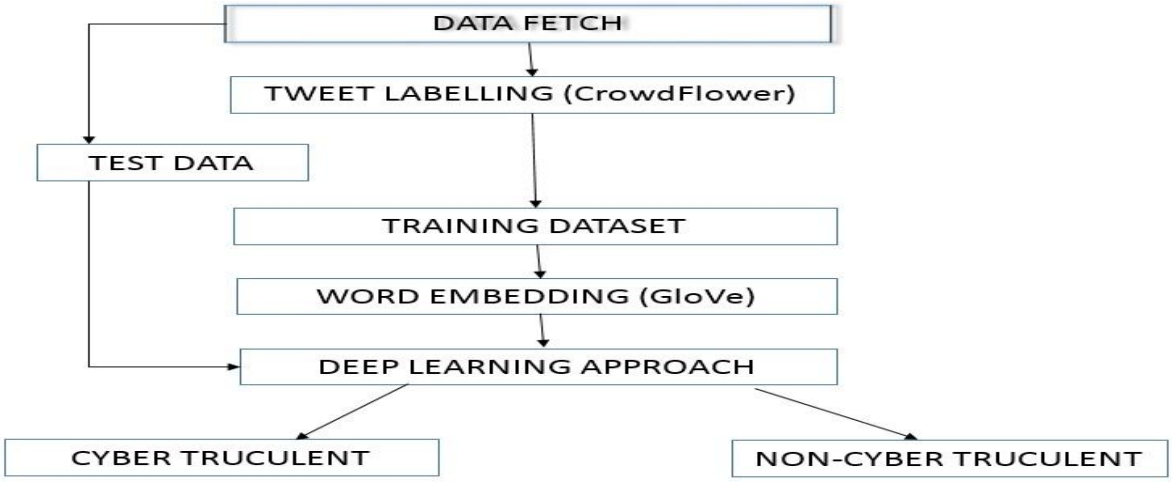

Figure 1: Developed Model

Data Collection

Data was originally gathered from social media platforms or cloud media by pulling reviews and information about reviews from websites and microblogging platforms (Twitter). However, it would be time-consuming and ineffective to manually copy and paste all the evaluations from the websites under consideration into a document; as a result, this procedure will be automated using a data scraping script that was built during the study. Twenty-four thousand, seven hundred and eighty-three (24,783) tweets with the class labels "bullying language," "non-bullying language," and "neither" made up the gathered dataset. In the study, tweets were categorized into the SE, SW, SS, NC, NE, and NW geopolitical zones. The notion of bullying or abusive behavior was used to categorize tweets as hate speech when they were tweets classified as neither featured either harsh or bullying language, according to the concept of bullying or truculent discourse. The dataset additionally includes a count column that displays the annotator's classification judgment for that tweet.

For downstream tasks, the dataset was divided into two parts: training (80% of the dataset was utilized for training, while 60% was used to validate the learned model), and performance testing (20%). The massive amount of duplicate records in the data collection is its primary shortcoming. According to analysis of the train and test sets, there are around 78 and 75 percent of duplicate data in each set, respectively. The train set's huge number of redundant records will hinder learning.

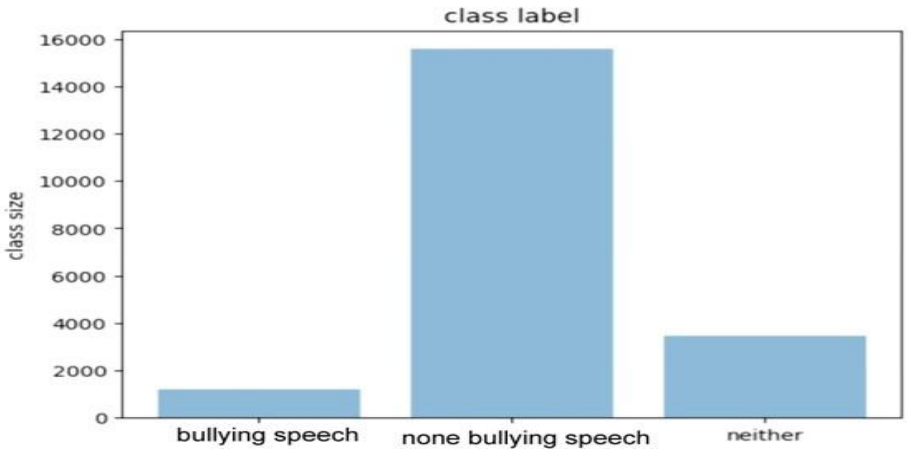

Figure 2: Class imbalance in the Dataset Sentiment Analysis

In contrast to other class labels, such as none bullying language and neither, the annotators were more consistent when it came to determining whether a tweet was a bullying speech or not. Table 1 presents details showing the hate speech of class label, which makes up only 5.7 percent of the entire dataset.

Table 1: Hate Speech Class Label

Class	Data Size (%)	Disagreement (Agreement)
Bullying Speech	1.430 Sample (5.77%)	87.4% (12.6%)
None Bullying Language	19.190 Sample (77.43%)	22.1% (77.9%)
Neither	4.163 Sample (16.79%)	28.8 %(71.2%)
Total	24.783 (100%)	

The study employed the tweeter comment dataset as it provides a varied range of sentences with varying numbers of opinion targets for training the maximum entropy classification model for target extraction and sentence type classification. It includes 14,492 sentences from a range of user sources that addressed pertinent sentiment categorization themes using the maximum entropy model, as shown below.

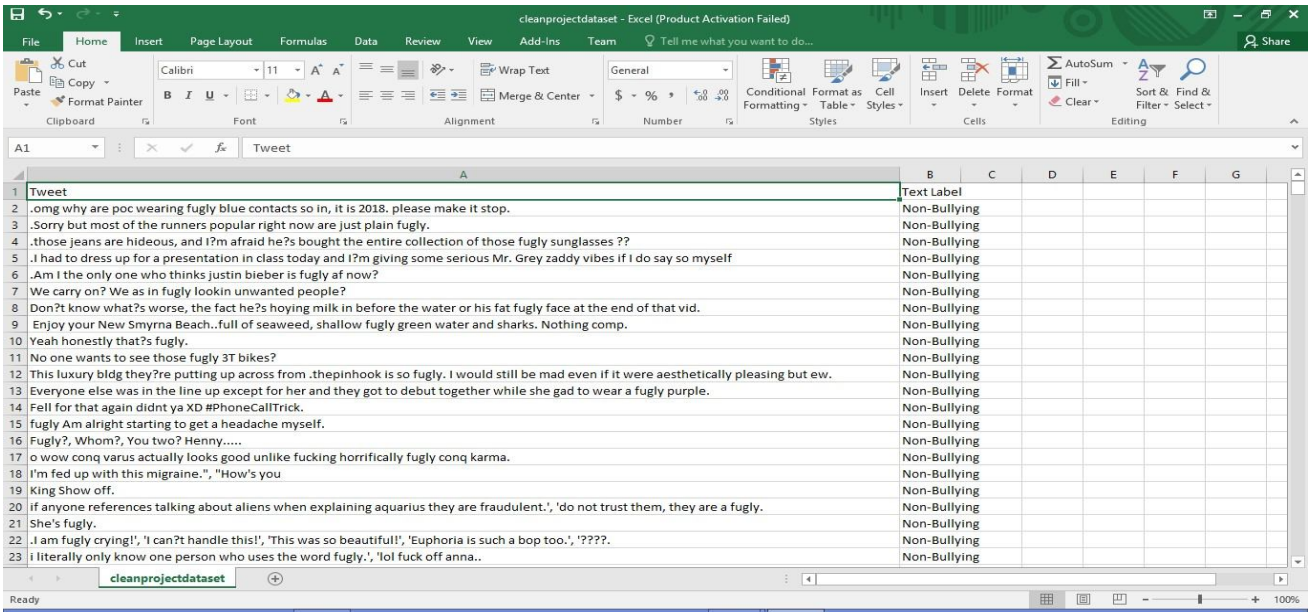

Fig 3: Cleaning of the Dataset for Classification

The research's training utilized a rule-based, subjective sentence detector. The hints are appropriate for our sentimental-based challenge since they have orientations to divisive, contentious, and opinionated themes. The study starts by identifying the semantic word properties that give subjective phrases their subjectivity in order to create the cyberbullying lexicon. We need to supplement the subjective semantic lexicon with corpus-generated vocabulary since the field of cyberbullying is so highly reliant on domain dependency and context-specific lexicon. We include verbs linked to bullying and dependency-type produced grammatical patterns related to the three indicated thematic areas into our vocabulary using bootstrapping and WordNet. The development of a cyberbullying detection tool with three degrees of "Bullying," "None Bullying," and "Neutral" based on the terminology

labeled paragraphs was gathered. The summarize process was carried out in four main steps:

Step 1: use a rule-learning approach to extract subjective sentences.

Step 2: Using subjective sentences identified in step 1 above, extract semantic and subjective word features was obtain.

Step 3: Using bootstrapping, we augment the lexicon in step 2 with noun patterns based on the semantic classes of religion, ethnicity and race and bullying-related verbs.

Step 4: Then build and test the classifier with the annotated corpus based on the features identified in Step 2 and Step 3.

The goal is to classify the document into a certain category (positive/neutral/negative, objective/subjective, etc.) using the contextual information in the document (unigrams, bigrams, other features within the text). Let w1,...,wm be the maximum number of words that can exist in a document that has to be optimized, using the conventional bag-of-words structure that is frequently used in natural language processing and information retrieval. The Lagrangian multipliers were first presented to help address optimization problems, and later an unconstrained dual problem was

used to estimate the lamda free variables 1,...,n via maximum likelihood estimation.

Estimating Lamda Parameters

Input : $Feature\ functions\ f_1, f_2, \cdots f_n;\ empirical\ distribution\ \tilde{p}(x, y)$

Output : $Optimal\ parameter\ values\ \Lambda_i^*;\ optimal\ model\ p^*$

1. $Start\ with\ \lambda_i = 0\ for\ all\ i \in \{1, 2, \cdots, n\}$
2. $Do\ for\ each\ i \in \{1, 2, \cdots, n\}$:
   a. $Let\ \Delta\lambda_i\ be\ the\ solution\ to$

$$\sum_{x,y} \tilde{p}(x)\, p(y \mid x)\, f_i(x, y) \exp\left(\Delta\lambda_i f^{\#}(x, y)\right) = \tilde{p}(f_i)$$

   where $f^{\#}(x, y) = \sum_{i=1}^{n} f_i(x, y)$

   b. $Update\ the\ value\ of\ \lambda_i\ according\ to: \lambda_i \leftarrow \lambda_i + \Delta\lambda_i$
3. $Go\ to\ step\ 2\ if\ not\ all\ the\ \lambda_i\ have\ converged$

The $f^{\#}(x, y)$ is the total number of features which are active for a particular (x, y) pair. If this number is constant for all documents then the $\Delta\lambda_i$ can be calculated in closed-form:

Deep Learning:

Sentences of variable lengths are inputted into the 1d-CNN, which generates fixed-length vectors. Word embedding are created for each word in the glossary of all input sentences prior to training. In a matrix M, all word embeddings are stacked. The words that make up the current training phrase are embedded in the input layer and obtained from M. The network is configured to accommodate phrases up to a certain length. Shorter sentences are padded with zero vectors, whereas longer sentences are eliminated. Then, over-fitting is managed via dropout regularization.
Multiple filters with various window sizes move on the word embedding to perform one-dimensional convolution in the convolution layer. Numerous sequences are formed as the filter progresses that capture the syntactic and semantic characteristics of the filtered n-gram.

To account for element-wise non-linearity, functions are included. The merge layer combines the outputs of various filters. A fully connected softmax layer produces the probability distribution across labels from various classes following another dropout procedure.
One of the most popular connectionism models for categorization is CNN. Models of connectionism emphasize learning from environmental cues and storing this knowledge as connections between neurons (figure 5). Some learning method adjusts the weights of a neural network in accordance with the training data.
In other words, the more the training data deviate from the real world, the harder it will be for the learning algorithm to adjust, and the poorer the classification results will be. separating opinionated statements into several categories based on the quantity

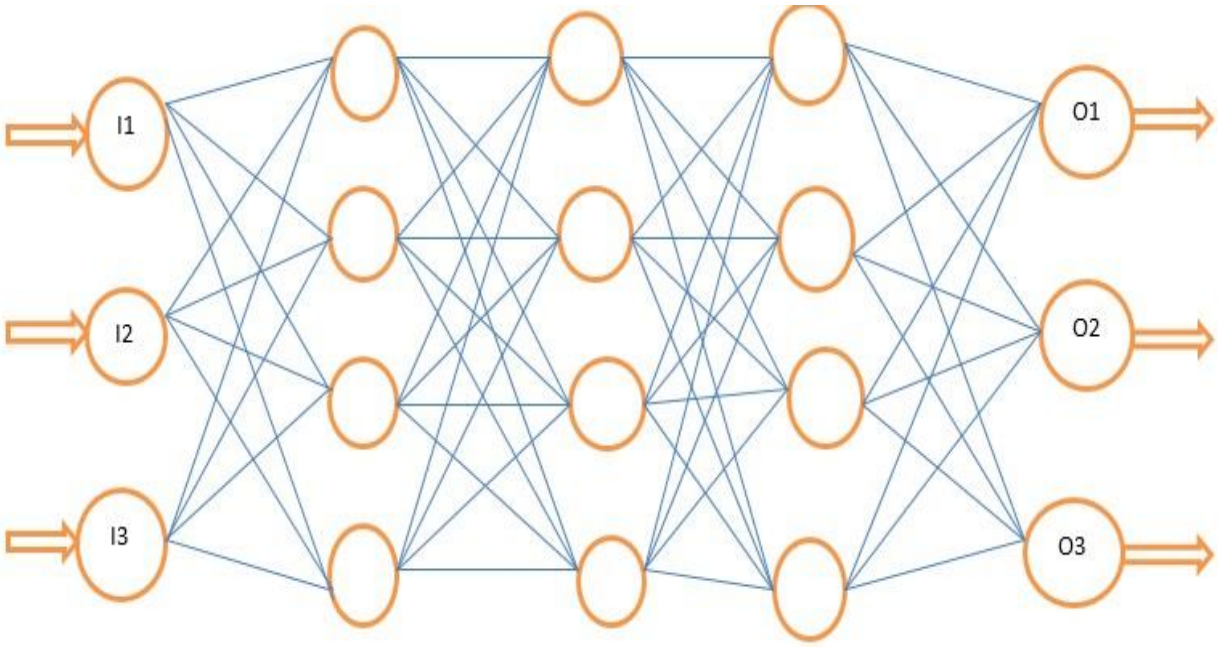

Figure 4: Convolutional Neural Network

RESULTS AND DISCUSSION

The following graphic illustrates the cyberbullying model for performance with Ngram, Trigrams recall, bullying precision, bullying recall, and accuracy. For performance evaluation, the most insightful aspects were attained. As can be shown in fig. 5 below, the created model was successful in forecasting the performance of cyberbullying.

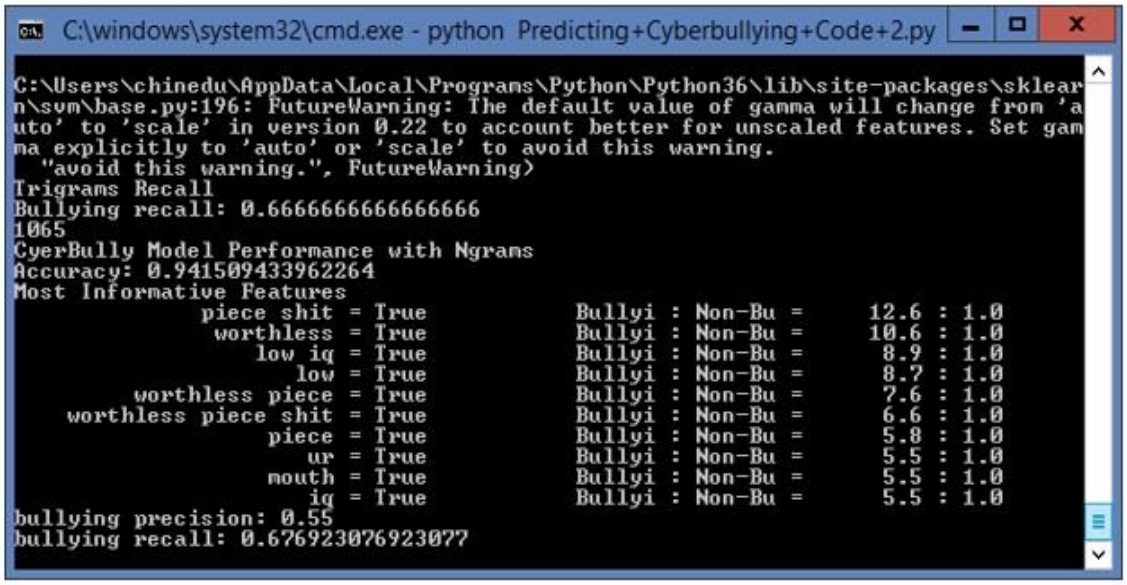

Fig 5: Predict the Cyberbullying Performance

Table 2 below provides the accuracy, f-score, and recall results: The accuracy, f-score, and recall results for the created model .

Table 2: Evaluation of Cyberbullying the Model

DATASET	ACCURACY	PRECISION	RECALL	F-SCORE
Unigram	96.3%	95.5%	73.7%	83.2%
Bigrams	93.8%	93.7%	67.6%	84.7%
Trigrams	88.2%	92.3%	85.7%	88.9%
Ngrams	94.2%	92.9%	67.7%	78.3%
Final	**97.9%**	**93.6%**	**73.7%**	**83.8%**

The representativeness of the dataset, which includes both training and test cases, is the main danger to this study's external validity. It goes without saying that the systems might behave differently during the test cases. The paper compares the suggested answer to a related piece of work by Banerjee et al. Figure 6 below displays the comparison's findings:

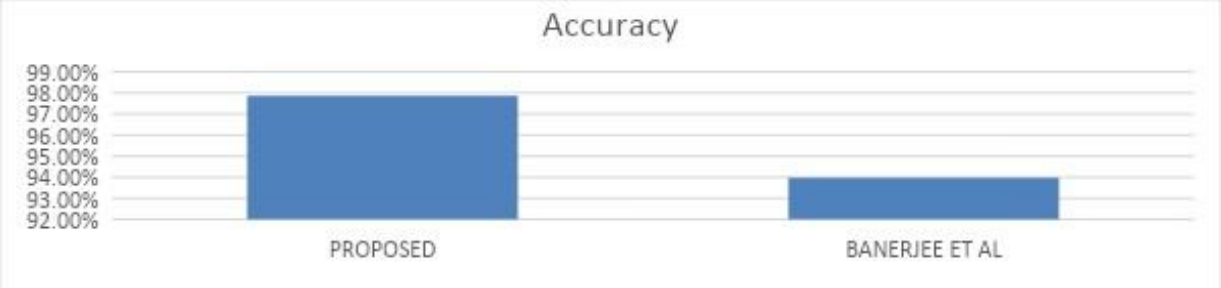

Fig 6: Graph of Accuracy Result

4. Conclusion

The data representations used in training the two baseline classifiers over a 50,000 enhanced tweets dataset were created using WordNet with character unigram, bigram, trigram, and n-grams recognized for identification of unusual words. The convolutional neural network and maximum entropy techniques were used to create the ensemble model. performed better in recognizing OOV words, different spellings, and smart language of cyberbullying in tweets than the individual models on their own, which meets the goal of the research effort while enhancing the recall accuracy of bullying. The model's unigram fared reasonably well, with 96.3 percent accuracy, while the bigram achieved the proposed model, which combines maximum entropy and convolutional neural network, outperformed the individual models, with an accuracy of 93.8 percent compared to the trigram's accuracy of 88.2 percent, the n-accuracy gram's of 94.2 percent, and the best-accuracy gram's of 97.9 percent. However, upon closer examination, it was discovered that the proposed model, which combines maximum entropy and convolutional neural network, performed better than the **individual models.**

REFERENCE

1. Ezeibe, ChristianHATE SPEECH AND ELECTORAL VIOLENCE IN NIGERIA. Conference: 10. Two-Day National Conference on The 2015 General Elections in Nigeria: The Real Issues at the Electoral Institute Complex, Independent National Electoral Commission Annex, Central Business District, . (2015). AbujaAt: Abuja.
2. Adibe, J. Fayose's advert: Offensive or hate speech? Adapted from a paper presented at a roundtable on hate speech organized by the Kukah Centre: Abuja, on 27 January. 2017.
3. Ravi, Kumar. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems. (2015). 89. 14-46.
4. Liu, B.. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. 10.1017/CBO9781139084789.
5. Appel, Orestes &Chiclana, Francisco & Carter, Jenny & Fujita, Hamido. A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level. Knowledge-Based Systems. (2016). 108. 10.1016/j.knosys.2016.05.040.
6. Baccianella, Stefano &Esuli, Andrea &Sebastiani, Fabrizio. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. Proceedings of LREC. 10.
7. Muhammad, N. Wiratunga, Robert Lothian . Contextual sentiment analysis for social media genres Computer Science Knowl. Based Syst. 2016.
8. Sunday Adeola AJAGBE , IfedotunRoseline IDOWU , John B. OLADOSU PhD and Ademola O. ADESINA (2020), Accuracy of machine learning models for mortality rate prediction in a Crime dataset International Journal of Information Processing and Communication (IJIPC) Vol. 10 No. 1&2 , pp. 150-160, ISSN 2645-2960; Print ISSN: 2141-3959
9. .Idowu, I. R., Adeniji O.d.., Elelu, S., &Adefisayo, T. O. Prediction of Breast Cancer Images Classification Using Bidirectional Long Short Term Memory and Two-Dimensional Convolutional Neural network. Transactions on Networks and Communicaitons, . (2021), 9(4). 29-38
10. Fernández Gavilanes, Milagros &Àlvarez-López, Tamara &Juncal-Martínez, Jonathan & Costa-Montenegro, Enrique & González-Castaño, Francisco. (2015). GTI: An Unsupervised Approach for Sentiment Analysis in Twitter. 533-538. 10.18653/v1/S15-2089.
11. Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1643–1654.
12. Meishan Zhang and Yue Zhang and Duy-Tin Vo (2015). Neural Networks for Open Domain Targeted Sentiment. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 612–621, Lisbon, Portugal, 17-21 September 2015.
13. Jindal, Nitin and Bing Liu. Identifying comparative sentences in text documents. in Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006). 2006a.
14. Park, MinJun& Yuan, Yulin. (2015). Linguistic Knowledge-driven Approach to Chinese Comparative Elements Extraction. 79-85. 10.18653/v1/W15-3114.
15. Adeniji OluwasholaDavid ,AkinolaOlaniyanEliais (2022). A Secured Text Encryption with Near Field Communication (NFC) using Huffman Compression. International Journal of Engineering and Applied Computer Science, Volume: 04, Issue: 02, March 2022 ISBN: 9780995707542.
16. Adeniji O. D, OsofisanAdenike. Route Optimization in MIPv6 Experimental Testbed for Network Mobility: Tradeoff Analysis and Evaluation. International Journal of Computer Science and Information Security (IJCSIS), (2020) Vol. 18, No. 5, pp 19-28.

17. Adeniji O.d., Olatunji O.O. “Zero Day Attack Prediction with Parameter Setting Using Bi Direction Recurrent Neural Network in Cyber Security”. International Journal of Computer Science and Information Security (IJCSIS), (2020) , Vol. 18, No. 3,111-118.
18. Adeniji Oluwashola David. DynamicFlowReductionSchemeUsingTwoTagsMulti protocolLabelSwitching(MPLS)inSoftwareDefineNetwork. International Journal of Emerging Trends in Engineering Research. (2022). Vol. 10, No. 3,141-147.
19. Liu, Mingya. The elastic nonveridicality property of indicative conditionals. Linguistics . (2019). Vanguard. 5. 10.1515/lingvan-2019-0007.
20. Davidov, Dmitry &Tsur, Oren &Rappoport, Ari.. Semi-supervised recognition of sarcastic sentences in twitter and Amazon. CoNLL 2010 - Fourteenth Conference on Computational Natural Language Learning, Proceedings of the Conference.